Cell Reports Methods, Volume *3*

Supplemental information

Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data

Junlin Xu, Jielin Xu, Yajie Meng, Changcheng Lu, Lijun Cai, Xiangxiang Zeng, Ruth Nussinov, and Feixiong Cheng

Supporting Information

Graph Embedding and Gaussian Mixture Variational Autoencoder Network for End-to-End Analysis of Single-Cell RNA-Sequencing Data

Xu et al., 2022

*Correspondence to: Feixiong Cheng, Ph.D.

Lerner Research Institute, Cleveland Clinic

Email: chengf@ccf.org

Tel: +1-216-4447654; Fax: +1-216-6361609

Supporting Information includes 6 Supplemental Figures (.pdf) and 6 Supplemental Tables (.xlsx).

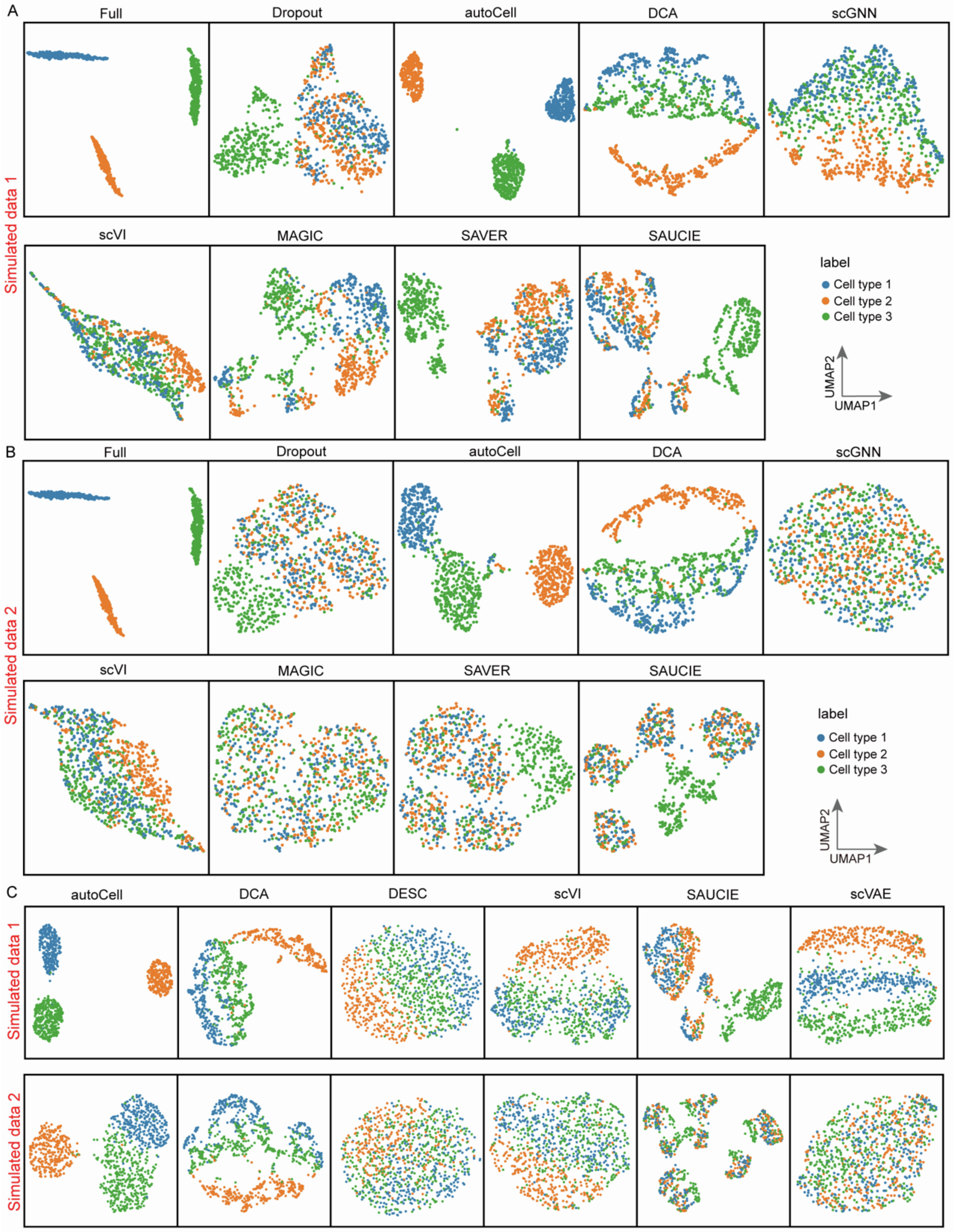


Figure S1. **UMAP visualization on simulated datasets, related to Figure 2 and Figure 4.**

(A) and **(B):** Comparison of the effect of imputation on cell type analysis. UMAP visualization using the dropout data and imputed ones by autoCell, DCA, scGNN, scVI, MAGIC, SAVER and SAUCIE for the Simulated data1 and Simulated data2, respectively. Each color represents one of the 3 cell types. **(C):** Feature embedding. UMAP visualization of the extracted features from autoCell, DCA, DESC, scVI, SAUCIE and scVAE of the Simulated dataset. For comparison, autoCell, DCA, DESC, scVI, SAUCIE and scVAE all performed dimension reduction to 10 dimensions before applying UMAP.

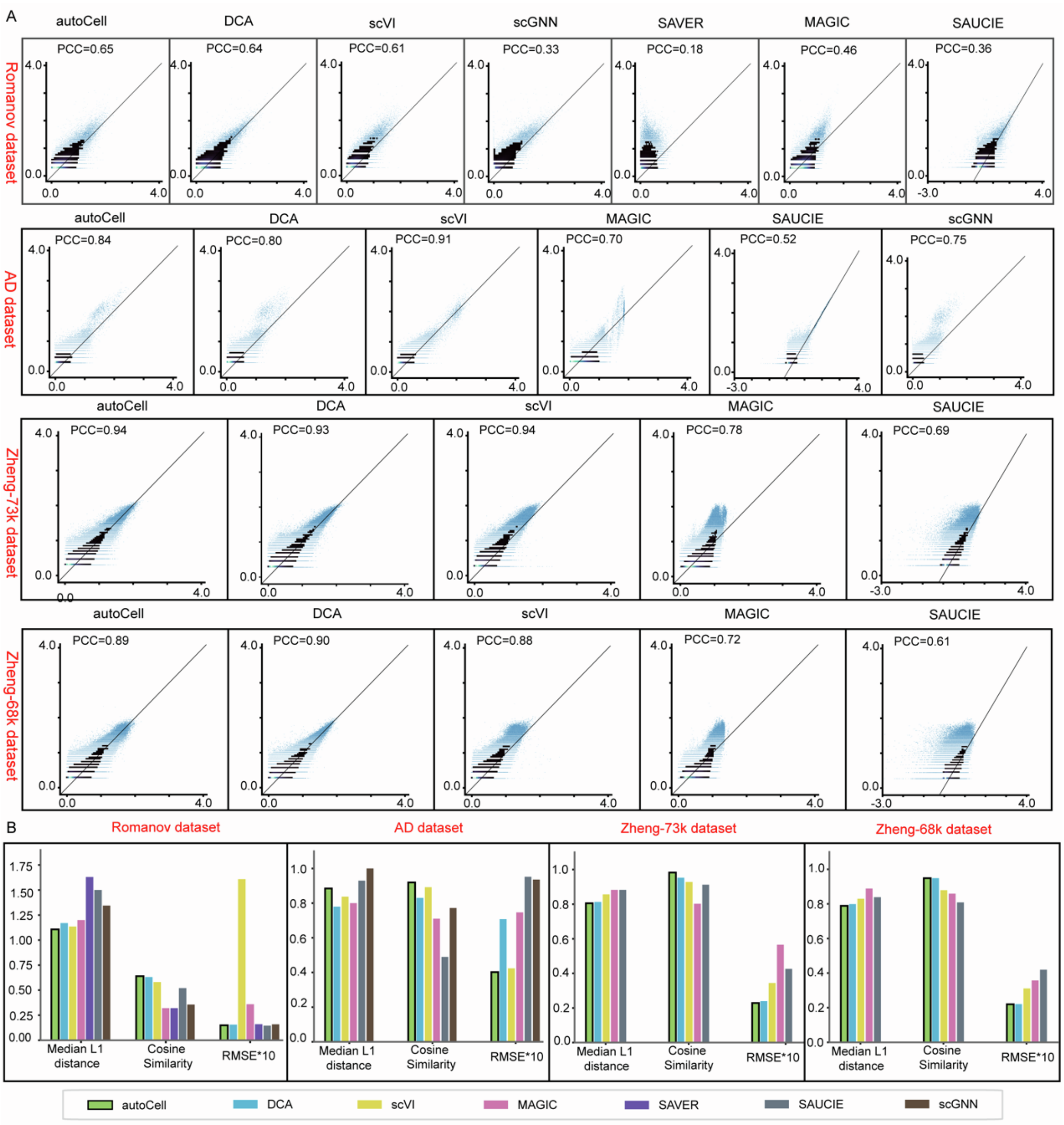
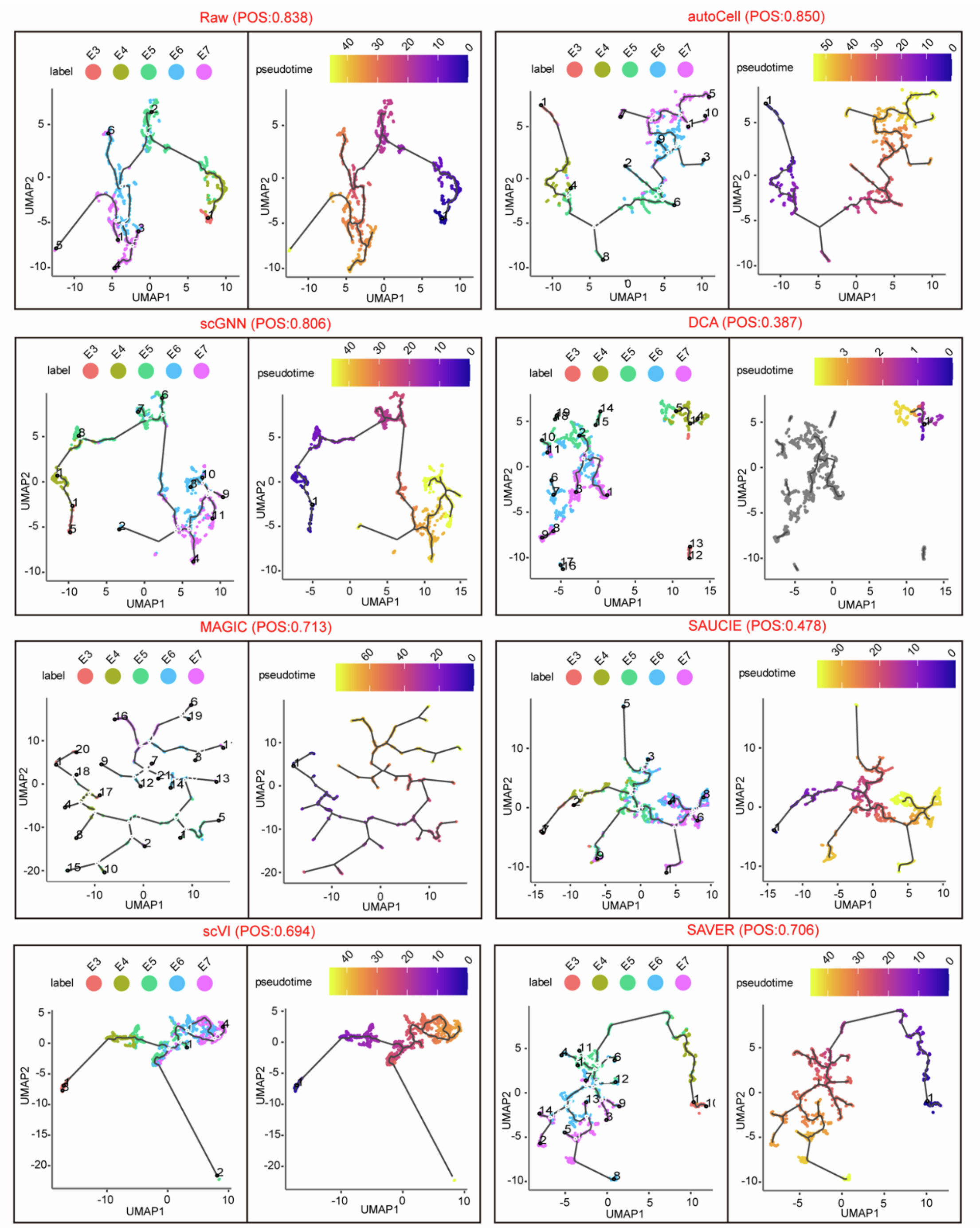


Figure S2. Performance comparison between autoCell and other state-of-the-art methods under 10% synthetic dropout rate on four real datasets, related to Figure 2.

(A): Density plots of imputed vs. original data masked. The x-axis corresponds to the imputed values, and the y-axis represents the true values of the masked data points. Each row is a different dataset, and each column is a different imputation method. **(B):** Comparison of the cosine similarity (higher is better), median L1 distance (lower is better), root-mean-square- error (RMSE) scores (lower is better), and Pearson correlation coefficient (PPC, higher is better) between autoCell and other imputation tools. Note: we cannot run SAVER and scGNN for some datasets (speed issues).



Raw (POS:0.838)

autoCell (POS:0.850)

scGNN (POS:0.806)

DCA (POS:0.387)

MAGIC (POS:0.713)

SAUCIE (POS:0.478)

scVI (POS:0.694)

SAVER (POS:0.706)

Figure S3. **autoCell improves pseudotime analysis in the human preimplantation embryonic development dataset, related to Figure 3.**

The results of Monocle3 estimated pseudotime using the raw data and imputed ones by autoCell, scGNN, DCA, MAGIC, SAUCIE, scVI and SAVER as input. POS: Pseudotime order score (higher is better).

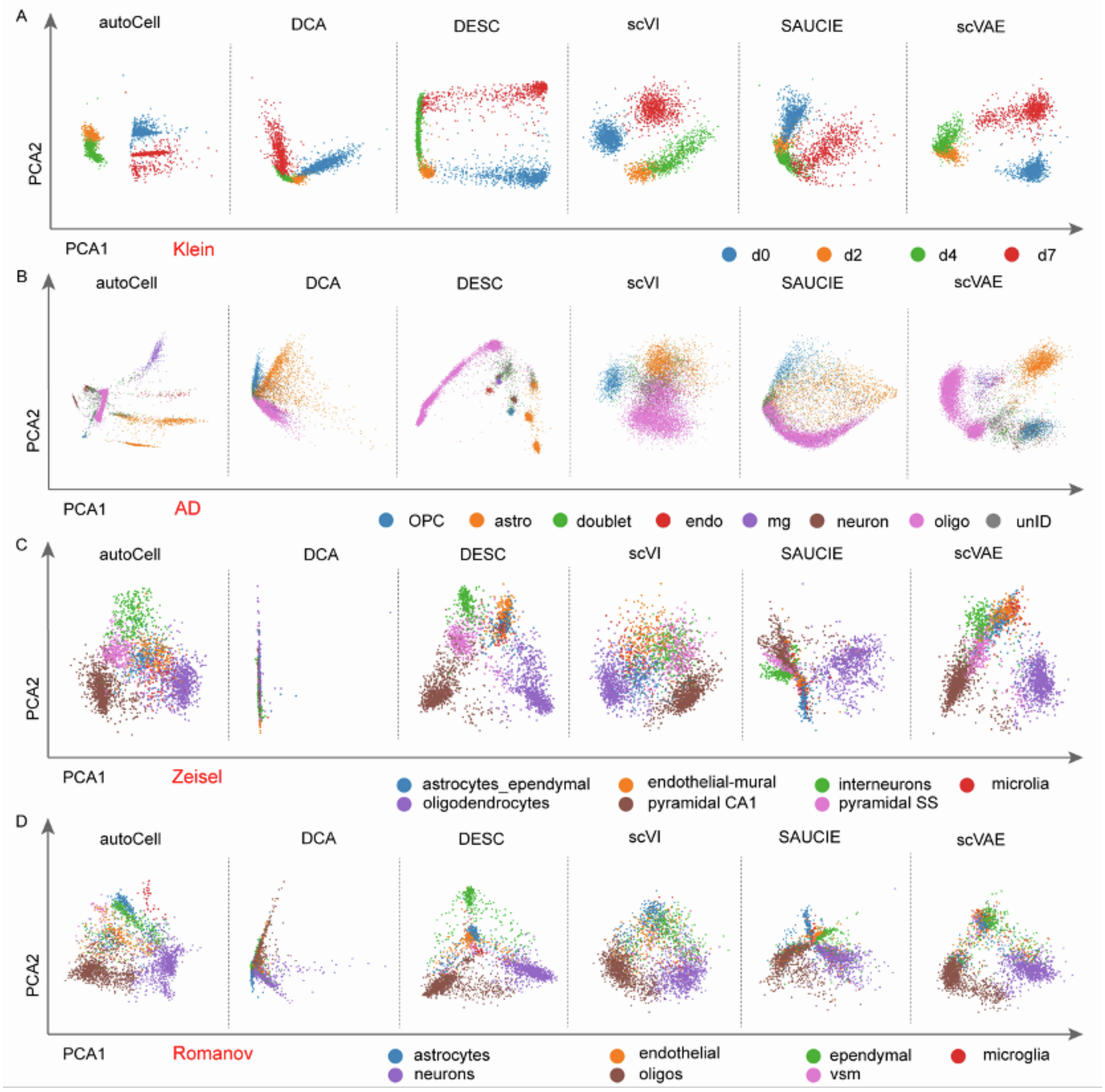


Figure S4. **PCA visualization of the extracted features using different approaches, related to Figure 4.**

PCA visualization of the extracted features from autoCell, DCA, DESC, scVI, SAUCIE and scVAE of 4 real datasets. For comparison, autoCell, DCA, DESC, scVI, SAUCIE and scVAE all performed dimension reduction to 10 dimensions before applying PCA.

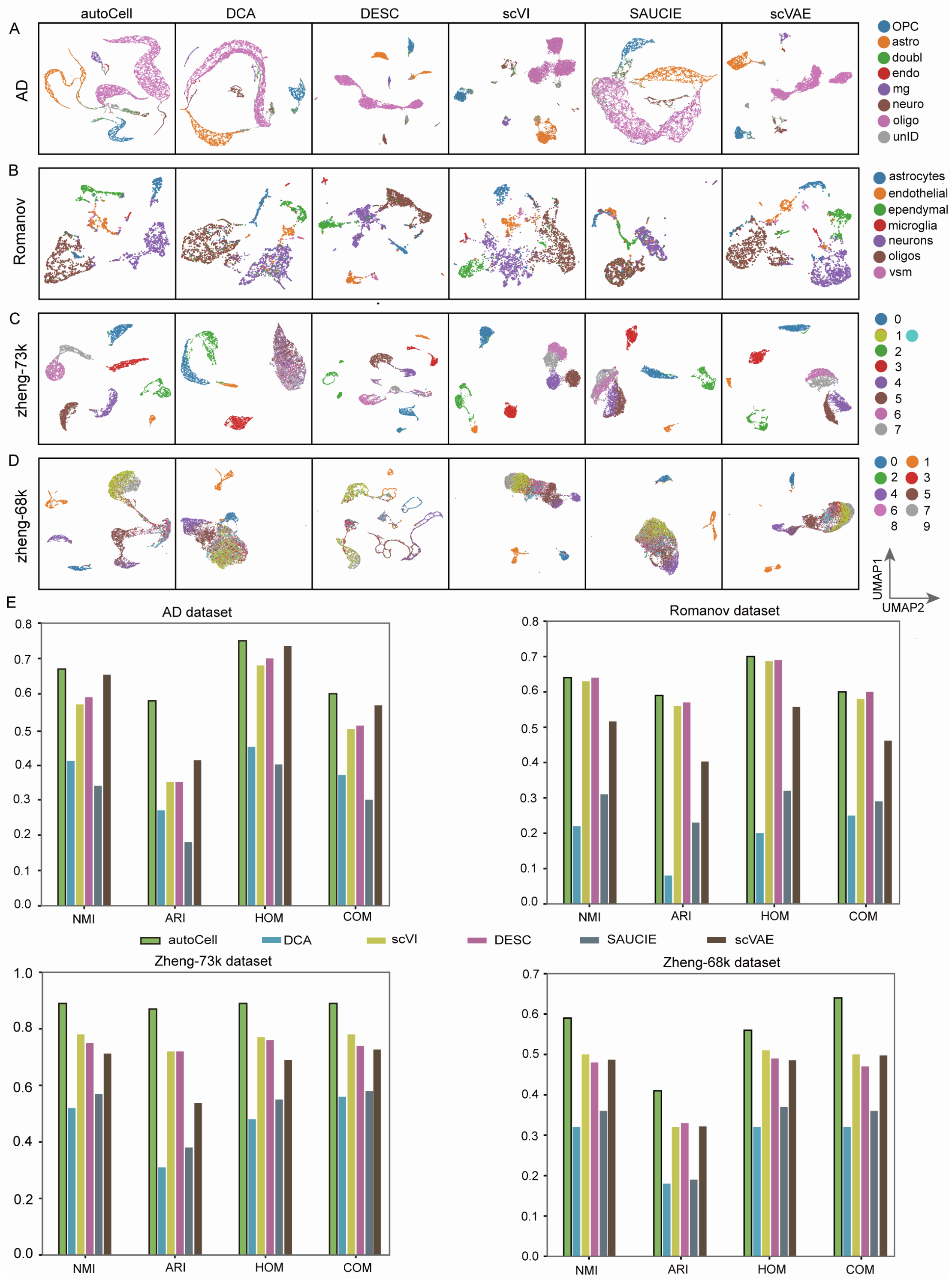


Figure S5. **UMAP visualization of the extracted features using different approaches, related to Figure 4.**

(A-D): UMAP visualization of the extracted features from autoCell, DCA, DESC, scVI, SAUCIE and scVAE of the different dataset. For comparison, autoCell, DCA, DESC, scVI, SAUCIE and scVAE all performed dimension reduction to 10 dimensions before applying UMAP. **(E)** Comparison on the effect of clustering on four datasets. Clustering accuracy was evaluated by applying K-means clustering on the extracted features to obtain cluster assignments. NMI: normalized mutual information. ARI: adjusted rand index. COM: completeness. HOM: homogeneity.

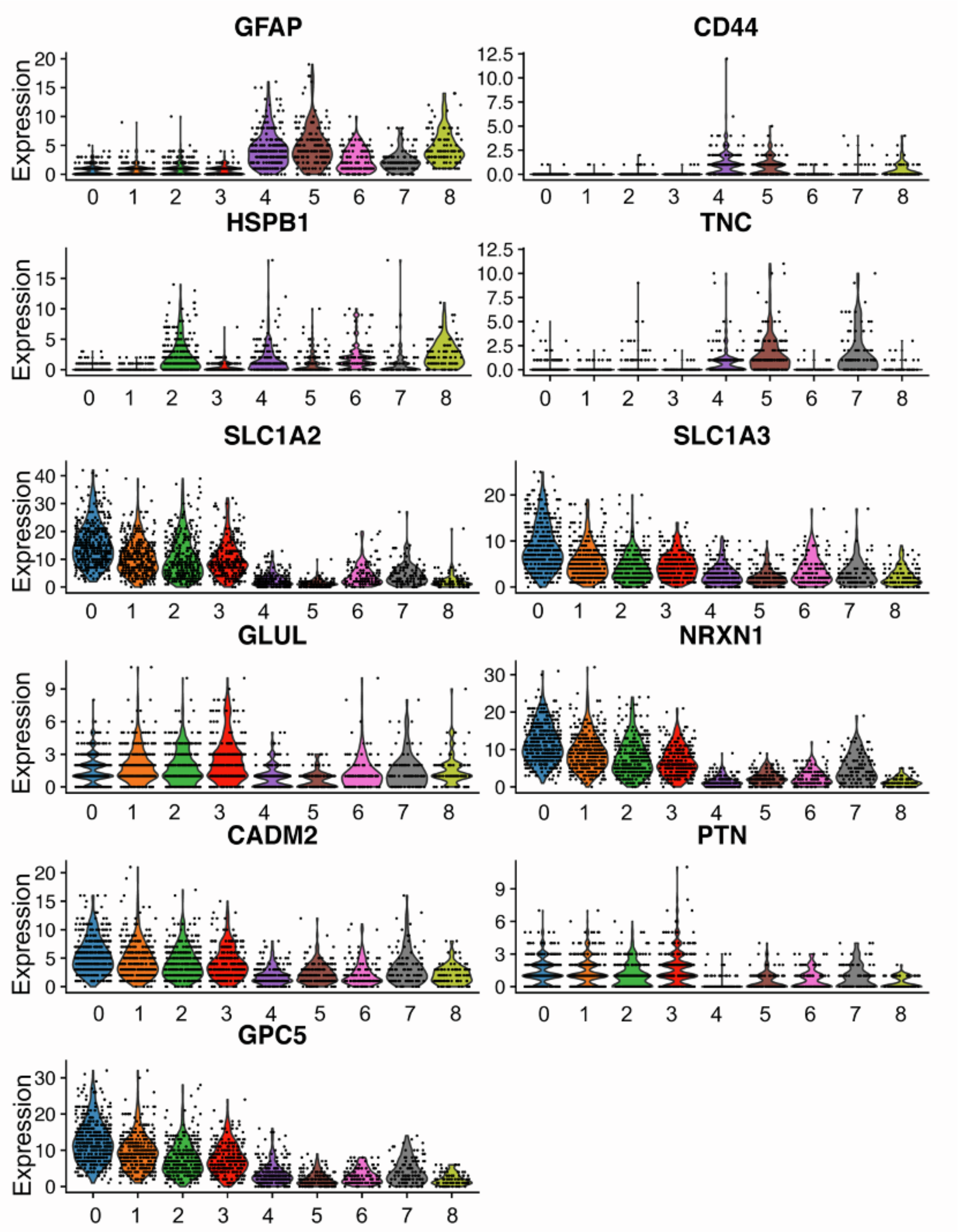


Figure S6. The expression levels of 11 identified DAA marker genes, related to Figure 5. Upregulated marker genes: GFAP, CD44, HSPB1 and TNS; and Downregulated marker genes: SLC1A2, SLC1A3, GLUL, NRXN1, CADM2, PTN and GPC5, across the 9 astrocyte subpopulations identified by autoCell.